A Deterministic Annealing Approach to Optimization of Zero-delay Source-Channel Codes

Mustafa S. Mehmetoglu, Emrah Akyol, Kenneth Rose
Dep. of Electrical & Computer Eng.
UC Santa Barbara, CA, US
Email: {mehmetoglu, eakyol, rose}@ece.ucsb.edu

Abstract*—This paper studies optimization of zero-delay source-channel codes, and specifically the problem of obtaining globally optimal transformations that map between the source space and the channel space, under a given transmission power constraint and for the mean square error distortion. Particularly, we focus on the setting where the decoder has access to side information, whose cost surface is known to be riddled with local minima. Prior work derived the necessary conditions for optimality of the encoder and decoder mappings, along with a greedy optimization algorithm that imposes these conditions iteratively, in conjunction with the heuristic "noisy channel relaxation" method to mitigate poor local minima. While noisy channel relaxation is arguably effective in simple settings, it fails to provide accurate global optimization results in more complicated settings including the decoder with side information as considered in this paper. We propose a global optimization algorithm based on the ideas of "deterministic annealing"- a non-convex optimization method, derived from information theoretic principles with analogies to statistical physics, and successfully employed in several problems including clustering, vector quantization and regression. We present comparative numerical results that show strict superiority of the proposed algorithm over greedy optimization methods as well as over the noisy channel relaxation.

I. Introduction

The zero delay source-channel coding problem has recently gained revived interest [1]–[5]. In this paper, we focus on numerical optimization of the zero-delay mappings. In prior work [6], a method, "noisy channel relaxation" (NCR) [7], [8] was employed to mitigate the poor local minima problem inherent to such optimization problems. While NCR is relatively successful in the point-to-point setting, it is insufficient to obtain precise results in more involved settings such as the decoder side information setting. In this paper, we incorporate a powerful non-convex optimization method, *deterministic annealing*, within a framework proposed in our prior work [6] to numerically obtain the globally optimal zero-delay mappings in the side information setting.

Deterministic annealing (DA) is a global optimization approach, based on information theoretic principles with analogies to statistical physics, that has been successfully used as a remedy to the problem of poor local minima in non-convex optimization problems, including clustering [9], vector quantization [10], regression [11] and more (see review in [12]). An important distinction between DA and other non-convex optimization tools such as NCR is that DA is independent of the initialization.

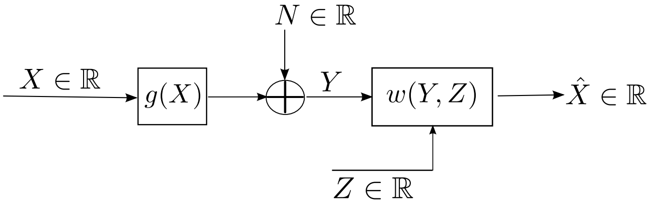

Fig. 1. The problem setting

This paper is organized as follows. In Section II, we present preliminaries and the problem definition. In Section III, we review prior work including the necessary conditions for optimality, and optimization aided by NCR. In Section IV, we describe the proposed algorithm. Numerical comparisons are presented in Section V and concluding remarks in Section VI.

II. Preliminaries and Problem Definition

Let $\mathbb{E}(\cdot)$, $\mathbb{P}(\cdot)$ and $\mathbb{R}$ denote the expectation and probability operators, and the set of real numbers, respectively. Let ∇ and ∇_x denote the gradient and partial gradient with respect to x, respectively. Let $f^{'}(x) = \frac{df(x)}{dx}$ denote the first order derivative of function $f(\cdot)$. The joint Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix R is denoted as $\mathcal{N}(\boldsymbol{\mu}, R)$. All the logarithms in the paper are natural logarithms and may in general be complex. The integrals are in general Lebesgue integrals. While we focus on scalar sources and noises, our results can easily be extended to vector spaces, albeit with more involved notations.

The problem setting is given in Figure 1, where source $X \in \mathbb{R}$ and side information $Z \in \mathbb{R}$ are drawn from joint density $f_{X,Z}(\cdot,\cdot)$. Z is available only to the decoder, while X is mapped to channel input by the encoding function $g : \mathbb{R} \to \mathbb{R}$ and transmitted over the channel whose additive noise $N \in \mathbb{R}$, with density $f_N(\cdot)$, is independent of X, Z. The received channel output $Y = g(X) + N$ is mapped to the estimate $\hat{X}$ by the decoding function $w : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. The problem is to find optimal mapping functions $g(\cdot), w(\cdot)$ that minimize the mean squared error (MSE) distortion

$$D = \mathbb{E}\{(X - \hat{X})^2\}, \tag{1}$$

subject to

$$P(g) = \mathbb{E}\{g^2(X)\} \leq P. \tag{2}$$

Although the problem we consider is delay limited, it is insightful to consider asymptotic bounds achievable at infinite delay. From Shannon's source and channel coding theorems, it is known that, asymptotically, the source can be compressed to $R(D)$ bits (per source sample) at distortion level D, and that C bits can be transmitted over the channel (per channel use) with arbitrarily low probability of error, where $R(D)$ is the source rate-distortion function, and C is the channel capacity, (see e.g. [13]). The asymptotically optimal coding scheme is the tandem combination of the optimal source and channel coding schemes, hence $R(D) \leq C$ must hold. By setting

$$R(D) = C, \tag{3}$$

one obtains a lower bound on the distortion of any source-channel coding scheme. The capacity of the additive white gaussian noise channel with variance σ_N^2 is given by

$$C = \frac{1}{2}\log(1 + \frac{P}{\sigma_N^2}), \tag{4}$$

where P is the transmission power constraint and σ_N^2 is the noise variance. For source coding with decoder side information, it has been established for Gaussians and MSE distortion that there is no rate loss due to the fact that the side information is unavailable to the encoder [14]. Hence, optimum performance theoretically attainable (OPTA) can be obtained by equating the conditional rate distortion function of the source (given the side information) to the channel capacity. The rate distortion function of X when Z serves as side information and $[X, Z] \sim \mathcal{N}(\mathbf{0}, R)$ where $R = \sigma_X^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ with $|\rho| \leq 1$ is:

$$R(D) = \max(0, \frac{1}{2}\log\frac{(1-\rho^2)\sigma_X^2}{D}), \tag{5}$$

We plug (5) and (4) in (3) to obtain OPTA

$$D_{OPTA} = \frac{(1-\rho^2)\sigma_X^2}{(1 + \frac{P}{\sigma_N^2})}. \tag{6}$$

III. Prior Work

Here, we summarize the relevant contributions of prior work, see [6] for more details.

A. Necessary Conditions for Optimality

Let the encoder $g(\cdot)$ be fixed. Then, the optimal decoder is the MSE estimator of X given $Z = z$ and $Y = y$:

$$w(y, z) = \mathbb{E}\{X|y, z\}. \tag{7}$$

Plugging the expressions for expectation, applying Bayes' rule and noting that $f_{Y|X}(y, x) = f_N[y - g(x)]$, the optimal decoder can be written, in terms of known quantities, as

$$w(y, z) = \frac{\int x\, f_{X,Z}(x, z)\, f_N[y - g(x)]\, \mathrm{d}x}{\int f_{X,Z}(x, z)\, f_N[y - g(x)]\, \mathrm{d}x}. \tag{8}$$

To derive the necessary condition for optimality of $g(\cdot)$, we consider the distortion functional

$$D[g, h] = \mathbb{E}\{(X - w(g(X) + N, Z))^2\}, \tag{9}$$

and construct the Lagrangian cost functional:

$$J[g, w] = D[g, w] + \lambda P[g]. \tag{10}$$

Now, let us assume the decoder $w(\cdot)$ is fixed. To obtain necessary conditions, we apply the standard method in variational calculus:

$$\nabla_g J[g, w] = 0,\ \forall x, \tag{11}$$

where

$$\nabla_g J[g, w] = \lambda f_X(x) g(x)$$
$$-\iint w'(g(x)+n, z)\, [x - w(g(x)+n, z)]\, f_N(n) f_{X,Z}(x, z) \mathrm{d}n\mathrm{d}z. \tag{12}$$

and $w'(\cdot,\cdot)$ denotes the derivative with respect to the first argument.

Remark 1: Note that the linear encoder and decoder mappings satisfy the necessary conditions for optimality in the Gaussian case. However, it is well known that linear mappings are highly suboptimal, see e.g. [6]. This fact illustrates the existence of poor local optima and the challenges facing algorithms based on these necessary conditions.

B. Greedy Algorithm

Iteratively alternating between the imposition of individual necessary conditions for optimality, will successively decrease the Lagrangian cost to a stationary point. Imposing the decoder optimality condition is straightforward, since it is expressed in closed form as functional of the encoding mapping $g(\cdot)$. The encoder optimality condition is not in closed form and we perform an appropriate steepest descent search. The encoder is updated as given below, where i is the iteration index and μ is the step size.

$$g_{i+1}(x) = g_i(x) - \mu\nabla_g J[g, w]. \tag{13}$$

At each iteration i, total cost decreases monotonically and iterations are kept until convergence.

There is no guarantee that an iterative descent algorithm of this type will converge to the globally optimal solution, in fact, simulations show severe issues of local optima. As a remedy, NCR method of [7], [8] was embedded in the iterative algorithm in [6], i.e., the algorithm was run for a very noisy channel (high Lagrangian parameter λ), and then gradually decrease λ while using the prior mapping solution as initial condition.

IV. Proposed Method

We recast the zero-delay source-channel coding problem as a regression problem optimizing for the encoding function within a given parametric class of functions. We restrict the discussion to piecewise regression functions which approximate the desired mappings by partitioning the space and matching a simple local model to each region. Such regression functions are determined by specifying two components: a space partition and a parametric local model per partition

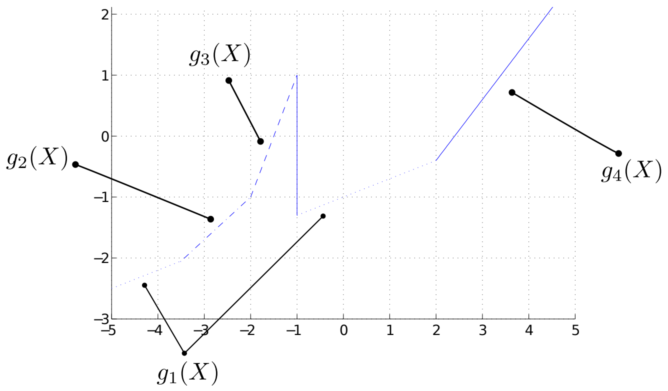

Fig. 2. An example encoder function consisting of affine local models, $K = 4$

cell (typically a simple model such as constant, linear, or Gaussian[1]).

DA introduces controlled randomization into the optimization process. The problem is recast as minimization of the expected cost subject to a constraint on the level of randomness as measured by the Shannon entropy of the system. The resulting Lagrangian functional can be viewed as the free energy of a corresponding physical system whose Lagrange parameter is the "temperature". The minimization is started at a high temperature (highly random mappings) where, in fact, the entropy is maximized and all points equally belong to the all partition cells (and effectively there is only one local model). This minimum is then tracked at successively lower temperatures (lower levels of entropy) as the system typically undergoes a sequence of phase transitions through which the model complexity (the number of distinct local models) grows. As the temperature approaches zero, the distortion and power terms dominate the Lagrangian cost and a hard (nonrandom) mapping is obtained.

We proceed to describe in more detail the proposed DA-based method.

A. Structured Encoder Functions

We consider the parametric functions (local models) $g_k(x) = f(x, \Lambda_k)$, for $k \in \{1, ..., K\}$, with the parameter sets Λ_k. These functions have a certain parametric form and each function is defined over a region denoted as $\mathbb{R}_k$. The overall encoder function is defined as $g(x) = g_k(x)$ for $x \in \mathbb{R}_k$. The parametric form is to be chosen appropriately depending on the involved distributions and the design constraints. Figure 2 shows an example structured encoder with affine local models of the form $g_k(x) = a_k x + b_k$.

B. Randomized Associations

We randomize the associations of the input points to the local models, or regions. We first define the probabilities

$$p_{K|X}(k|x) \triangleq \mathbb{P}\{x \in \mathbb{R}_k\}, \quad \forall k, x. \tag{14}$$

Note that $\sum\limits_{k=1}^{K} p_{K|X}(k|x) = 1 \; \forall x$. Next, we rewrite (1) as

$$D = \sum_{k=1}^{K} \int\limits_{\mathbb{R}} D_k(x) p_X(x) p_{K|X}(k|x) \mathrm{d}x, \tag{15}$$

where $D_k(x)$ is the contribution to the distortion, when point x is associated with region k. It is given by

$$D_k(x) = \int\limits_{\mathbb{R}} d(x, w(g_k(x) + n, z)) p_N(n) p_{Z|X}(z|x) \mathrm{d}z \mathrm{d}n. \tag{16}$$

The power constraint in (2) is rewritten as

$$P = \sum_{k=1}^{K} \int\limits_{\mathbb{R}} g_k^2(x) p_X(x) p_{K|X}(k|x) \mathrm{d}x. \tag{17}$$

The cost function to minimize is

$$J = D + \lambda P \tag{18}$$

$$= \sum_{k=1}^{K} \int\limits_{\mathbb{R}} J_k(x) p_X(x) p_{K|X}(k|x) \mathrm{d}x, \tag{19}$$

where

$$J_k(x) \triangleq D_k(x) + \lambda g_k^2(x) \quad \forall k. \tag{20}$$

We now restate the problem as that of minimizing J over the local model parameters and association probabilities. Note that, given the local models, the association probabilities that minimize (19) will implement 'hard' associations, that is, every point is associated with probability one to the region that contributes the minimum cost to (20). Therefore, by randomizing the encoder we generalize the search space but preserve the same global minimum as the original problem.

C. Entropy Constraint

As we noted above, the direct optimization of the association probabilities will result in 'hard' probabilities. However, in order to avoid poor local optima we impose and control the level of randomness, i.e. we introduce a constraint on the randomness of the encoder, which we measure by the Shannon entropy. The total entropy of the encoder is given by $H(X, K) = H(X) + H(K|X)$ and since $H(X)$ is constant (determined by the source) we define $H \triangleq H(K|X)$ where

$$H(K|X) = -\int\limits_{\mathbb{R}} p_X(x) \sum_{k=1}^{K} p_{K|X}(k|x) \log(p_{K|X}(k|x)) \mathrm{d}x. \tag{21}$$

Remark 2: It is important to note that the approach is generalizable to the "mass-constrained" variant of DA [15], where entropy maximization is effectively replaced by minimization of the mutual information $I(K; X)$. Such generalization offers additional optimization advantages (see [15]), as well as a useful and direct link to rate-distortion theory (see [16] for analysis of these connections, as well as DA for rate-distortion function computation). The corresponding "mass-constrained"

[1] In this paper, we use only affine models, however it is straightforward to include other models within the optimization framework.

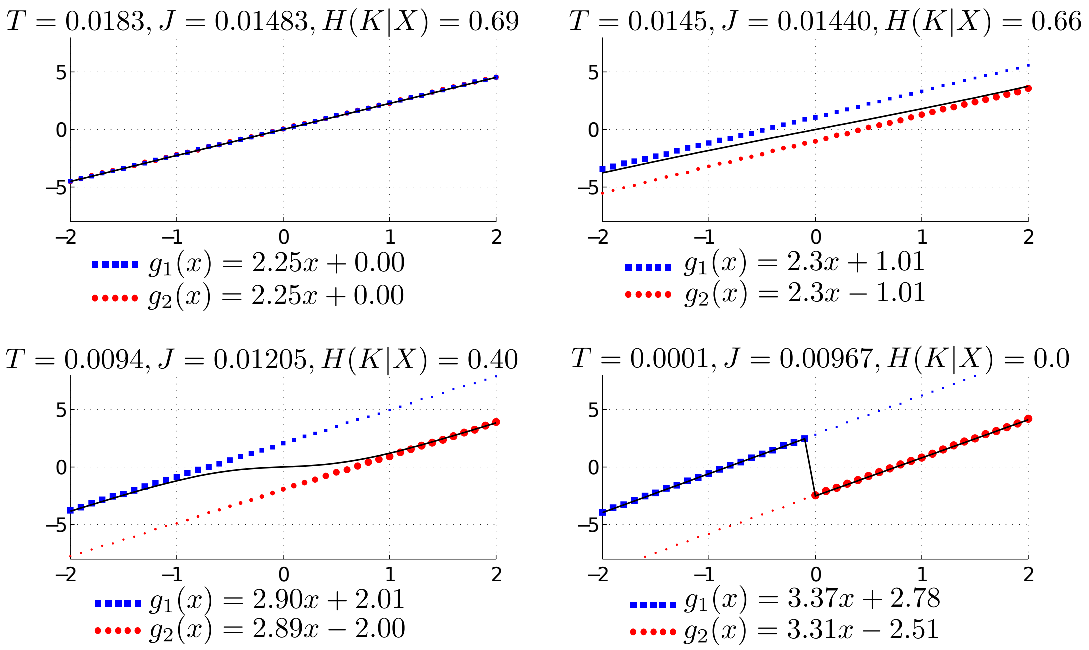

Fig. 3. The evolution of the encoder in the algorithm. The two models are shown by dotted lines and the size of a dot gives the probability association at that input point to that local model. The black line represents the averaged encoder, $K = 2$.

extension for the current problem is a work in progress and is outside of the scope of this paper.

Accordingly we construct the Lagrangian

$$F = J - TH, \tag{22}$$

to be minimized, with T (temperature) being the Lagrange multiplier associated with the entropy constraint. Note that for large T, the minimum F is achieved by maximizing the entropy. At lower values of T, randomness is traded for reduction in J. In the limit $T = 0$, minimizing F corresponds to minimizing J directly, which produces a deterministic encoder. Therefore, we start at a high value of T and gradually lower it while minimizing F at each step.

We present an example of the method in Figure 3 with two local models[2]. When T is large, the local models are coincident. As we lower T, the system goes through a bifurcation point (referred as "phase transition" in statistical physics) where the two local models split from each other to decrease F. The corresponding value of T is referred as the first "critical temperature". Further phase transitions can be obtained by keeping a duplicate for each local model at every temperature. The duplicates will merge at every iteration until a critical temperature is reached, and will split at a phase transition.

The pseudocode of the method is given in Algorithm 1.

D. Update Equations

The optimum local model parameters cannot be obtained in closed form, hence we perform gradient descent search. The gradient with respect to any local parameter θ_k from a set Λ_k can be obtained as

$$\frac{\partial F}{\partial \theta_k} = \frac{\partial J}{\partial \theta_k} = \int\limits_x p_X(x) p_{K|X}(k|x) \frac{\partial [D_k(x) + \lambda g_k^2(x)]}{\partial \theta_k} \mathrm{d}x. \tag{23}$$

For the affine model, θ_k denotes a_k and b_k.

The association probabilities that minimize F are derived in a straightforward fashion as the Gibbs distribution

$$p_{K|X}(k|x) = \frac{e^{-[D_k(x) + \lambda g_k^2(x)]/T}}{\sum\limits_{k=1}^{K} e^{-[D_k(x) + \lambda g_k^2(x)]/T}} \quad \forall x. \tag{24}$$

Remark 3: As expected, (24) results in uniform associations for large T and "hard" (binary) associations for $T = 0$.

The optimum decoder given the encoder can be derived by plugging $p(y|x,z) = \sum\limits_{k=1}^{K} p_N(y - g_k(x)) p_{K|X}(k|x)$ in (8).

Algorithm 1 The outline of the proposed algorithm

```
Initialize: High T, single region (K=1)
while $H(K|X) > H_{min}(K|X)$ do
    Duplicate (if $K < K_{max}$) and perturb local models
    while $cost_{i+1} < cost_i$ do
        update the local model parameters
        update $p_{K|X}(k|x)$ $\forall k, x$
        update $w(y, z)$
    end while
    Check if regions have split
    Set $T = \alpha T$                                ▷ e.g. $\alpha = 0.95$
end while
```

[2] The example is run for jointly Gaussian source and side information and Gaussian noise.

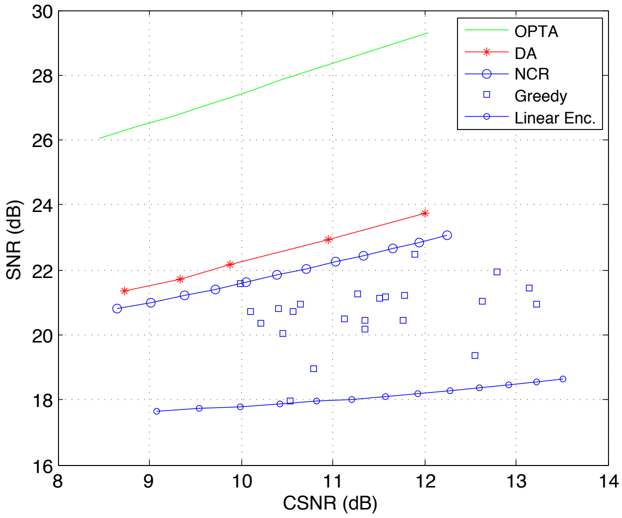

Fig. 4. The performance comparison of the proposed method with greedy optimization, the noisy relaxation (NCR), and the linear mappings.

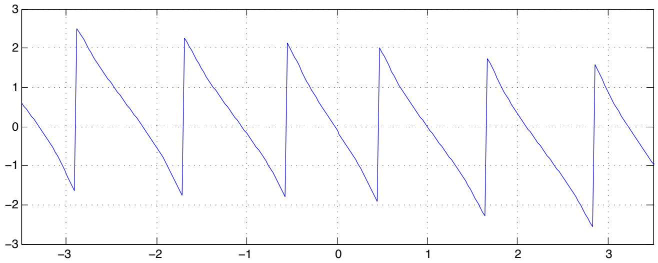

Fig. 5. An example mapping for correlation coefficient $\rho = 0.99$ at $CSNR = 10.98$, $SNR = 23.2$.

V. Experimental Results

The comparative performance results are given in Figure 4 for jointly Gaussian X and Z with $\rho = 0.99$ as described in Section II, and Gaussian noise with unit variance. The best results of the NCR method out of multiple runs, and various results from the greedy method are presented in Figure 4. Note that the proposed method is independent of the initialization and only run once, whereas the results of greedy approach and NCR heavily depend on initialization, as can be seen from various points obtained by the greedy approach. We also present the performance of OPTA as benchmark while noting that it is asymptotic and requires infinite delay. The performance of linear encoder and decoder is plotted as well, since it is also a local minimum (see Remark 1).

An example mapping from the same setting is also given in Figure 5. Interestingly, as noted before (see e.g., [5], [6]) the analog mapping captures the central characteristic observed in digital Wyner-Ziv mappings, in the sense of many-to-one mappings, where multiple source intervals are mapped to the same channel interval, which will potentially be resolved by the decoder given the side information. However, we see differences between the mappings obtained by NCR (see e.g., [5], [6]) and ones by the proposed DA based method, e.g, the linear trend of the encoding mapping, that yield significant performance improvement as shown in Figure 4. Such differences are difficult to obtain and very important for the design of parametric mappings, see e.g., [4].

VI. Conclusions

In this paper, we studied the problem of finding globally optimum encoder and decoder pairs in zero delay source-channel coding, focusing on the setting where a side information is available to the decoder. Since the cost surface is riddled with locally optimum points, we developed a method based on the deterministic annealing approach to obtain globally optimum points. The numerical results show superiority of the proposed algorithm over greedy optimization methods and as well as the previously adopted approach, i.e., NCR. As future work, we will investigate adopting our DA approach to obtain optimal mappings in more complicated network settings as well as well-known open control problems such as the Witsenhausen's counterexample.

References

[1] F. Hekland, P. Floor, and T. Ramstad, "Shannon-kotel-nikov mappings in joint source-channel coding," *IEEE Trans. on Comm.*, vol. 57, no. 1, pp. 94–105, 2009.

[2] Y. Hu, J. Garcia-Frias, and M. Lamarca, "Analog joint source-channel coding using non-linear curves and mmse decoding," *IEEE Transactions on Communications*, vol. 59, no. 11, pp. 3016–3026, 2011.

[3] V. Vaishampayan and S. Costa, "Curves on a sphere, shift-map dynamics, and error control for continuous alphabet sources," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1658–1672, 2003.

[4] X. Chen and E. Tuncel, "Zero-delay joint source-channel coding for the Gaussian Wyner-Ziv problem," in *Proc. IEEE Int. Symp. on Inf. Theory*, 2011, pp. 2929–2933.

[5] J. Karlsson and M. Skoglund, "Optimized low delay source channel relay mappings," *IEEE Transactions on Communications*, vol. 58, no. 5, pp. 1397–1404, 2010.

[6] E. Akyol, K. Rose, and T. Ramstad, "Optimized analog mappings for distributed source channel coding," in *Proceedings of IEEE Data Compression Conference*, 2010.

[7] S. Gadkari and K. Rose, "Robust vector quantizer design by noisy channel relaxation," *IEEE Transactions on Communications*, vol. 47, no. 8, pp. 1113–1116, 1999.

[8] P. Knagenhjelm, "A recursive design method for robust vector quantization," in *Proc. Int. Conf. Signal Processing Applications and Technology*, 1992, pp. 948–954.

[9] K. Rose, E. Gurewitz, and G. C. Fox, "Statistical mechanics and phase transitions in clustering," *Physical review letters*, vol. 65, no. 8, pp. 945–948, 1990.

[10] ——, "Vector quantization by deterministic annealing," *IEEE Transactions on Information Theory*, vol. 38, no. 4, pp. 1249–1257, 1992.

[11] A. Rao, D. Miller, K. Rose, and A. Gersho, "A deterministic annealing approach for parsimonious design of piecewise regression models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 2, pp. 159–173, 1999.

[12] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[13] T. Cover and J. Thomas, *Elements of information theory*. J.Wiley New York, 1991.

[14] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[15] K. Rose, E. Gurewitz, and G. Fox, "Constrained clustering as an optimization method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 8, pp. 785–794, 1993.

[16] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1939–1952, 1994.